INDUSTRIAL CLASSIFICATION IN THE CANADIAN CENSUS OF MANUFACTURES: Automated Verification Using Product Data

by

John S. Crysdale

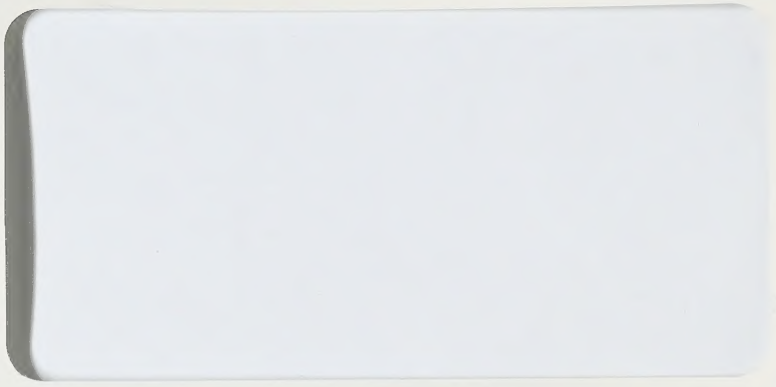
No. 20

Statistics Canada

Analytical Studies Branch

Research Paper Series

Canada



INDUSTRIAL CLASSIFICATION IN THE CANADIAN CENSUS OF MANUFACTURES: Automated Verification Using Product Data

by

John S. Crysdale

No. 20

Business and Labour Market Analysis Group
Statistics Canada

January 1989

An earlier version of this paper is published in the Statistical Journal of the U.N. Economic Commission for Europe. Volume 5, number 4, December 1988, 377-392.

Aussi disponible en français.

Industrial Classification In the Canadian Census of Manufactures: Automated Verification Using Product Data

John S. Crysdale

Abstract

Proper industrial classification of establishments is fundamental to the achievement of useful industry statistics. This paper describes an automated procedure used by Statistics Canada to verify industry coding in the Census of Manufactures. The edit also serves as a check on the accuracy of the detailed product data upon which the industry calculation is based. Some exceptions to the simple algorithm are reviewed, along with its impact on measures of industrial homogeneity. A number of suggestions are made as to how the assignment of industry codes can be further standardized and automated.

Acknowledgments

Harley Potter actively encouraged both this paper and the implementation of the methodology it describes. Thanks are also due to Statistics Canada personnel who provided detailed comments, especially to Ken Young, Chief, Analysis and Development Section, Industry Division, as well as to Shaila Nijhowne, Director, Standards Division and John McVey, Small Business and Special Surveys. The Analytical Studies Branch seminar provided a useful sounding board for an earlier version of this document. The methodology described in this paper was developed by the author while in the Analysis and Development Section of Industry Division. Those of the Census of Manufactures who actually lived through all these details helped immeasurably.

Key Words

SIC, Standard Industrial Classification, automated industry classification, computerized industry classification, Census of Manufactures, ICC, Industrial Commodity Classification, Harmonized Commodity Description and Coding System, economic statistics, manufacturing data, industry homogeneity, data quality, automated edit and imputation.

Industrial Classification In the Canadian Census of Manufactures: Automated Verification Using Product Data

John S. Crysdale

Systems of Industrial Classification

An Ancient Chinese Classification of Animals: Animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's brush, (l) others, (m) those that have just broken a flower vase, and (n) those that resemble flies from a distance. [1]

There are many systems of industrial classification--just as there are many systems of classification in general. They range from the rough and ready perceptions of market participants to the more formal structures of statistical agencies. While the results of industrial classification by these various methods will not necessarily coincide, what the different taxonomies do have in common is that each represents an effort to organize the relevant universe into groups that are similar in some useful way.

Statistics Canada now assigns industry codes according to the Standard Industrial Classification of 1980. This covers the entire universe of businesses in Canada--except for the so-called underground economy--and hierarchically divides and

subdivides that universe into groups of increasingly similar units.

What is meant by similar? Traditionally, an industrial classification involves similarity in terms of activities. But, presumably it could refer to any of a number of characteristics including size, location, country of control, or environmental impact.

Within manufacturing, similarity in terms of activities means, in practice, similarity in terms of materials used, market served (commercial/household, male/female), process of production, or end use of the product. In establishing the Standard Industrial Classification of 1980, it was not considered useful to employ just one of these criteria throughout all the manufacturing industries. Sometimes a given criterion would be appropriate, sometimes not. For example, end use can be difficult to assess in cases where manufacturers sell intermediate goods.

The nature and degree of similarity can change over time. In developing the 1980 version of the Standard Industrial Classification, there was a change in the scope of manufacturing itself and, within that, narrower activity definitions led to a number of 1970-based industries being subdivided. It is conceivable that in the next revision of the Standard Industrial Classification there will be more emphasis on chief component material. This would result from the implementation of the Harmonized Commodity Description and Coding System.

It might be desirable to simultaneously employ more than one industrial classification system, each with its own definition of similarity. Researchers could then use the

classification which most closely suited the purpose at hand. With increased automation in the coding process, multiple systems are entirely possible. [2] However, there could be problems if the resultant data were to be intermixed. In addition, confidentiality considerations would make release difficult.

Assigning Industry Codes to Manufacturing Units

In order to classify manufacturing units, a number of preliminaries must be addressed.

First, the classification unit must be determined. In the Canadian Census of Manufactures that unit is the establishment--often referred to as a plant, factory or mill. Once the unit is classified, then that classification applies to all its activities--even those that, conceptually, do not seem to belong. For example, if a fish processing plant freezes blueberries as a sideline, then that is included in the fish processing industry.

The alternative to classifying the entire unit to one industry is to prorate the data. This could involve arbitrary and complex machinations to make the figures conform with Census definitions of shipment values, and to generate the corresponding input data. Putting everything produced by an establishment into one industry has been the traditional solution.

Second, there must be a system for classifying products. At Statistics Canada, commodities reported to the Census of Manufactures are classified according to the Industrial Commodity Classification (ICC). As earlier noted, this is soon to be replaced by an extension of the Harmonized

Commodity Description and Coding System (which is being adopted to enhance the comparability of data from different sources and countries).

Third, there must be a linkage between the systems of industry and product classification. The two are related to one another by virtue of the fact that activities define industry classes. Such defining activities, and by extension the associated products, are then said to be primary to the industry in question. [3]

Operationally, if a product is primary to an industry, then that is the industry to which a plant producing only that product would be assigned. In other words, if fish canning is primary to the fish processing industry, then that is the industry to which a plant engaged solely in producing canned fish would be assigned.

All products not primary to a given industry but reported by establishments classified to it are said to be secondary. Where secondary activity exists, the industry is incompletely specialized. That corresponds to undercoverage of the defining activities of another industry.

Fourth, there must be an acceptable measure of the activity inherent in producing each commodity. Value added is preferred but is generally not feasible to calculate at the commodity level. Instead, establishments covered by the Census of Manufactures report commodity shipments or production. If those data cannot be substituted for value added, albeit with some loss of precision, it would be very difficult to apply a consistent set of industry coding rules across all units. Instead, assignments would be based on rather subjective nature of business enquiries.

Finally, the coding sequence must be determined. Apart from assignment on the basis of value added to one of the divisions of the Standard Industrial Classification, the approach is to code directly to the four-digit level. The major group and industry group codes are the first two and three digits of the four-digit industry class code--and follow automatically. Coding from the top down would be an alternate way of proceeding but would require three separate calculations and is not the procedure followed at Statistics Canada.

Given these preliminaries, and given that detailed product data do in fact exist, industrial classification is straightforward. There are two basic cases to consider: single- and multi-commodity plants. For the single-commodity producer, the technically correct classification is to the one item's primary industry. For the more common case, the multi-commodity plant, outputs are first grouped by primary industry and the highest-valued group determines the classification. This is the industry in which the unit is most specialized. It is not necessarily the industry to which the highest-valued single commodity belongs.

With a few exceptions, this simple plurality-based algorithm yields the correct industry assignment.

Verification of Industry Coding

(i) Background

It should be understood that, whenever a census or survey is conducted and the completed questionnaires are returned, the data are subject to a number of errors. For example,

required fields may not have been completed and, where answers are provided, they may be incorrect. This incorrectness may involve entry on the wrong line, lack of precision in applying concepts, or failure to provide accurate totals. Data capture introduces further potential for error. In any statistical agency, editing is a necessary function.

Editing also extends to the most basic assignment, namely the industrial classification of statistical units. In the Canadian Census of Manufactures, during the processing of data for 1983, the simple industry coding algorithm described earlier was incorporated into an automated editing tool known as the Questionably-Coded program. In the inevitable vernacular, this came to be referred to as The Q-Coded.

Prior to the introduction of the Questionably-Coded routine, industry code checks were done manually, on a somewhat subjective basis, and if and when the subject matter officer saw fit.

From 1983 on, the automated edit was performed every two weeks upon plants whose data were well advanced in the editing process. The results were distributed to the officers, and it was up to them to take action. As part of the final preparation for each industry's data release, the program was executed one last time to verify all establishments classified to the industry in question.

Over time, the Questionably-Coded program has incorporated extra features. One of these involved calculating proxies for value added to help determine whether the unit should be considered a manufacturer at all. (As noted, shipments rather than value added are used to classify establishments

within manufacturing.) Other features involve listing three full years of data, and generating reports which allow management to more closely monitor the officers' responses to edit messages.

In contrast to manual efforts, the automated industry code edit is almost exhaustive. It covers all establishments reporting commodity data. Such details are required on both long and short form questionnaires. [4] In 1985, these covered 28,655 establishments--of an active manufacturing, logging and forestry universe of 43,183 records--and accounted for over 96% of manufacturing shipments. [5]

For Census year 1986, the Questionably-Coded routine was embedded in the overall manufacturing editing package. This means that investigating problems is less time-consuming since the relevant documentation does not need to be separately retrieved from the files. This, in turn, reduces the possibility that delayed response can allow an industry transfer to be precluded by the prior publication of the recipient industry. Integration of this edit into the overall package also means that the computer can prevent release of industry-level data prior to completing the edit for all constituent plants. The integrity of the results is still dependent upon the quality of manual intervention.

(ii) Operational Aspects

Using the algorithm described above, the Questionably-Coded program calculates an industry code and compares the result with the code already on file. If the two differ, there is a problem. The difficulty may lie with the previously assigned code or with the computerized data upon which the calculated code is based. In either case, a page of data is generated

which includes a list of commodities shipped by the offending establishment along with values and primary industry links.

An example of such a printout is reproduced in the Appendix. Both the establishment and its data are imaginary and are intended only for illustrative purposes.

Without very detailed industry knowledge, editors or officers can generally do little to determine the nature of the problem by just inspecting the printout. The original questionnaire needs to be examined, trade indexes consulted and possibly a call made to the respondent. If such efforts are not made then the system as it now operates could be totally undermined. Under such circumstances, industry codes could be indiscriminately altered to suit whatever classification is dictated by the product data--an undesirable situation if that information has not been verified. Similarly, commodity codes could be manipulated to be consistent with assigned industry classes.

There are a number of possible outcomes to the enquiries made following receipt of the Questionably-Coded printout. In about half the cases the data will be corrected in some fashion. The most obvious possibility is that the establishment's assigned industry code is no longer appropriate. The officer will transfer it. In Census years 1983, 1984 and 1985, there were a total of 2,762 transfers, on a 1980-basis, within manufacturing. [6] Many of these changes involved fine-tuning, but a large number were more substantial and involved changes of major group or industry group. 1,404 transferred at the 4-digit level, 693 changed at the 3-digit level, and 665 switched major groups.

Another possibility is that incorrect commodity values may be on file as a result of errors in reporting or in the data capture process. The insertion of an inadvertent digit, for example, can cause a secondary activity to balloon in importance.

A third possibility is that the commodity coding itself could be in error. The respondent might have entered data for one commodity on a line designated for another. When this is corrected, and the industry code recalculated, the assigned code would be consistent with the data.

Finally, the linkages between commodities and industries require occasional amendment. Such questions are the responsibility of the subject matter officers. However, after a year or two's use in a verification context, changes to these primary industry links are infrequent.

Exceptions to the Simple Algorithm

The other half of the cases encountered by the verification routine are judged to be anomalies and are deliberately left as such. For these, an override code is tagged to the record so the Q-Coded routine will not reopen the case until the next year.

Table #1 summarizes the use of overrides in 1984 and 1985. The overrides are described below.

(i) Override #1: Commodity Code Problems

Some commodity codes do not effectively differentiate between similar products and are therefore too imprecise to be useful for industry coding. For example, at one time there existed

just one code for flower pots and it was treated as being primary to the plastics industries. A plant producing only clay flower pots would be calculated as belonging to the plastics industries. Assignment to the clay products area would be more reasonable. The interim solution was to override the calculated code. Later, because values justified it, two new commodity codes were created by adding a digit to the previous code.

After such a split, the original code is still not useful for industry coding as it continues to embed subclasses primary to different industries. The difference is that such cases--those having explicit subclasses--can be detected by computer, flagged, and made primary to no industry. Use of such imprecise codes has been virtually eliminated.

(ii) Override #2: Stability Problems

Another reason to permit technically incorrect coding is to avoid fluctuations in the data that do not reflect reality for the industry being measured. For example, suppose a plant with total shipments of $100 produces goods primary to Industries A and B, 49% and 51%, respectively. If total shipments in the next year were the same but the percentage split was reversed, then shipments primary to Industry A would have increased by $2, from $49 to $51. Transferring this plant to Industry A would magnify that slight change into a $100 increase for the industry. There would be a corresponding fiftyfold impact on Industry B. If marginal changes like this were to persist for at least two years, then the transfer would be made. With more substantial shifts, one might be tempted to make an immediate transfer.

Table #1

Application of Overrides & Coverage of the Industry Code Edit 1984-85

	Number of Statistical Units		Value of Manufacturing Shipments ($C billions)	
	1984	1985	1984	1985
Override Applied:				
#1 Commodity Codes	330	126	3.4	5.2
#2 Stability	383	642	2.1	5.8
#3 Timeliness	551	317	1.8	0.8
#4 Miscellaneous	69	49	0.9	0.3
#5 Insignificance	-	767	-	0.6
Total	1,333	1,901	8.2	12.8
Coverage of this Edit	26,514	28,655	228.7	245.9
Universe	42,977	43,183	235.7	254.3

In the U.S. Annual Survey of Manufactures a similar approach is followed in the industry coding of larger establishments. The precise rules are referred to as a resistance formula.

There are two special cases which may not be obvious. One involves units that produce products primary to a number of industries. Slight shifts in the mix can lead to continual transfer. To avoid this instability, assignment may be made to one of the miscellaneous industries. A second case, to be discussed later, involves the Machine Shop Industry.

(iii) Override #3: Timeliness Problems

Industry results of the Census of Manufactures are not all published at one time, but rather over a period of months as editing and analysis are completed. This means that an establishment may be left in an incorrect industry because the proper industry has already been published (or is very close). This happens more frequently towards the end of the processing cycle. It would not happen in systems where there is a preliminary data release followed by simultaneous release of final figures for all industries.

(iv) Override #4: Miscellaneous Problems

This code is used to deal with the various unusual circumstances that are inevitably found in actual data processing.

On occasion, deliberate miscoding has been employed for confidentiality purposes. Almost all establishments in the Canadian northern territories have been assigned to SIC 3999, Other Manufacturing Industries Not Elsewhere Classified.

This allows release of provincial industry totals with no loss of publishable northern detail. In another instance, moving a large establishment to a small, stable industry would have effectively released its confidential data.

Another category involves cases where coverage rather than specialization is deemed more important. If a producer is a major player in what is, technically, a secondary activity and has a plurality of output primary to, say, a residual industry class, then the eventual assignment may be based on coverage.

Coverage considerations might also apply to units that have a substantial amount of manufacturing activity, but a plurality in merchandising. An example might be a multinational which imports finished goods to supplement its domestic production. Ultimately, the respondent may be pressed to file two reports which, if successful, would suggest that application of the establishment concept was previously incorrect.

Other explanations given for using the miscellaneous override have included cases where it was not possible to verify accuracy of the data, as for example when the firm has gone out of business. Sometimes, the calculation of value added will be judged by the officer to produce an inappropriate result. For example, if an agricultural cooperative buys, processes and sells food products, it may set its input prices after the fact so as to distribute all its profits. That would eliminate manufacturing value added.

(v) Override #5: Insignificant Impact on Data

It was believed that the timeliness code sometimes masked the real reason for leaving the industry assignment unchanged.

Accordingly, an additional override category was introduced during the processing of Census year 1985. This covers cases which, in the officer's judgement, have an insignificant impact on the industry data and which, apart from a possible adjustment to the following year's mailing list, are deemed not worth investigating or correcting.

Further Exceptions

Some preliminary adjustments are required before the simple algorithm can be used to verify two unusual groups within the Standard Industrial Classification of 1980.

The Printing, Publishing and Allied Industries include classes defined in terms of joint production. Even a dollar's worth of publishing will turn a printer into a combined printer and publisher. For example, an establishment doing nothing but printing continuous forms is assigned to SIC 2811, Business Forms Printing Industry; but if the same plant also publishes school textbooks, the continuous forms are treated as primary to SIC 2849, Other Combined Publishing and Printing Industries, and the textbooks, otherwise primary to SIC 2831, Book Publishing Industry, become primary to SIC 2841, Newspaper, Magazine and Periodical (Combined Publishing and Printing) Industry. This adjustment requires testing for the joint presence of printing and publishing, altering the primary designation of individual products and then applying the simple algorithm as before.

In the Paper and Allied Products Industries, there are instances where the degree of vertical integration is taken into account. For example, cut newsprint is primary to two industries: SIC 2712, Newsprint Industry, and SIC 2799, Other

Converted Paper Products Industries Not Elsewhere Classified. In the former case, it is the result of an integrated process starting with wood chips. In the latter, it involves cutting or converting a large roll of newsprint which is the output of other establishments. For industry coding purposes, the presence of wood chips among the inputs is taken to indicate a fully integrated plant. Consequently, this adjustment requires examining input data and altering the primary designation of individual outputs before proceeding with the usual plurality-based calculation.

Measures of Industry Homogeneity [7]

Industries, it has been seen, comprise plants which are engaged in similar activities. More precisely, industries comprise plants which are mainly engaged in similar activities: the presence of secondary activities introduces an element of dissimilarity. [8]

There are two standard and complementary measures of industry homogeneity.

The specialization ratio, usually expressed as a percentage, indicates the extent to which activities reported by an industry are primary to it. By way of example, if $80 million of an industry's shipments of $100 million is primary to the industry, its specialization ratio is 80%.

The coverage ratio is the percentage of the value of an industry's defining activities that are actually reported by it. If all other industries account for a further $40 million of the above industry's primary outputs, its coverage ratio is 66.7%.

Measured homogeneity is almost always certain to be less than 100%. If just one establishment has secondary activity, specialization for that industry and coverage for its counterpart will be affected. In designing industrial classification systems, efforts to achieve higher levels of measured homogeneity must be restrained by conceptual considerations. Including the freezing of blueberries in the definition of fish processing might increase measured homogeneity but perhaps at the expense of conceptual uniformity. If the objective is to maximize empirical homogeneity at all costs, one need define only one industry--the economy.

Nevertheless, these ratios are useful for a variety of purposes. They are used (along with such criteria as economic significance) in devising new industry classes and in helping users interpret data such as concentration ratios. They can also be used to indicate the extent to which proper industry assignment has increased internal homogeneity of the data.

In the latter application it is, ironically, not always the case that more accurate coding increases measured homogeneity at the industry level. If an implicated establishment is unspecialized relative to the other units in the receiving industry, and specialized relative to the units in the sending industry, a transfer will decrease specialization for both industry classes. However, for the universe within which these are transferred, the weight shift accompanying this transfer will cause specialization--a weighted average--to rise.

Impact of the Industry Code Edit on Homogeneity Measures

In order to estimate the magnitude of data improvement arising from the introduction and use of the Questionably-Coded routine, it would be desirable to have before and after years in which nothing had changed other than that the industry code edit had been implemented. But there are a few difficulties in this regard.

First, there is a problem for comparability in that the introduction of this edit in 1983 coincided with (and assisted in) the transition from the Standard Industrial Classification of 1970 to the 1980-based version. The data were not collected on the same basis. However, as part of the conversion process, 1980-based industry codes were retroactively appended to each record for data-years 1981 and 1982. For the 1982 records, there was even some use of this edit. Consequently, use of that year's data as a starting point for measuring quality gains may lead to some understatement.

Second, the algorithm for calculating ratios has changed somewhat as have some of the primary industry assignments. To the extent that such modifications reflect a change in actual practice, using the revised methodology on all years' data may exaggerate the increase in homogeneity.

Another way to view the impact of the Questionably-Coded program would be to determine how homogeneity ratios differ within a given year from what they would have been if the industry code assigned to each unit had remained unchanged from the mailing of the questionnaire. The mailing is based on a file which contains the previous year's designations revised to reflect any subsequent births or transfers.

Industry coding is important on the mail-out file as it determines which one of a multitude of questionnaires is sent to each unit.

There are also a few difficulties with this approach to measuring the impact of the Questionably-Coded routine. It neglects the fact that the program handles non-industry code problems and overstates gains where transfers would have been made anyway. There is no audit trail to help untangle these effects. In addition, the before and after periods are not entirely comparable since cases are added during the processing year which did not appear on that year's mailing list. (Such additions are handled here by using the eventual classification as the mail-out code.)

Yet another way of examining the impact of this routine is to compare homogeneity ratios based on actual and calculated industry codes. Potential homogeneity refers to the specialization and coverage which would result if all plants were to be mechanically assigned to the industry calculated by the verification program. This is the theoretical maximum given the existing product mix and the primary industry linkages.

In Table #2, homogeneity ratios are presented for the period 1981 to 1985 using mail-out, actual and calculated 1980-based industry codes.

All homogeneity measures are rising, but the major observation is that, after 1983, the transition year, those based on mail-out and actual codes are very close.

Table #2

Homogeneity Ratios, Canadian Census of Manufactures*

	1981	1982	1983	1984	1985
Specialization					
° mail-out	n.a.	n.a.	86.7	89.0	88.1
° actual	86.5	88.3	89.2	89.9	89.2
° potential	90.3	90.9	91.0	91.6	91.9
Coverage					
° mail-out	n.a.	n.a.	88.0	90.2	90.0
° actual	88.3	89.7	90.5	91.0	91.1
° potential	90.7	91.3	91.4	91.8	91.8

* These ratios are presented as weighted averages at the all manufacturing level. The 1981 and 1982 mail-outs were conducted based upon the 1970 version of the Standard Industrial Classification; 1980-based data are unavailable.

In 1983, there was much fine-tuning of the new 1980-based codes and a clean-up of accumulated cases. In that year there were 1,537 transfers effected after the mail-out--i.e. during the processing cycle. The improvement between the mail-out and the actual data reflects this activity.

In 1984 and 1985, transfers continued to improve the data--although the results were less dramatic. In those years, a total of 1,407 transfers occurred during the processing cycle. 705 of these cases involved long forms, and of these, 456 moved outside the original 3-digit industry group. Under existing processing techniques, all transfers--but particularly these 456--involved considerable human and computer costs and some loss of timeliness.

The reduced impact of industry transfers after 1983 may suggest that transfers implemented during the processing cycle be restricted to cases exceeding a critical size. Different conclusions could be drawn where there is a greater degree of automation and where there are both preliminary and final data releases.

Limitations of this Procedure for Automated Coding

It would be desirable to be able to fully automate the process of industrial classification in the Canadian Census of Manufactures. To do so would require that all units can be accurately coded according to a series of prespecified rules. At present, this is impeded by a number of considerations, outlined in what follows, which collectively mean that the automated industry code edit is somewhat constrained even as a device to review manual coding and to

flag cases for further manual intervention. These considerations also restrict the potential to use the routine to experiment with revisions to industry or commodity classifications, or to retroactively recode units to extend industry-level time series.

(a) Data-Related Considerations

(i) There are insufficient data to allow the industry code edit to completely verify the divisional assignment of each manufacturing unit. The existing data allow an approximate calculation to be undertaken as to whether merchandising would be more appropriate. However, other potential assignments could include mining, construction or agriculture. More data are needed--but are acquired at the expense of increased response burden.

(ii) One third of active manufacturing units are not directly verified. These include head offices, ancillary units (such as garages and warehouses), and establishments whose product data are combined with and reported by a related establishment. However, all these units are verified indirectly inasmuch as their classification depends upon that of the establishments with which they are linked. The remaining cases that are not directly verified are units for which data are derived from administrative sources (coded by nature of business enquiries) and any small establishments for which commodity data have been estimated by computer. All the cases described above account for less than 4% of manufacturing shipments.

(iii) Accurate industry coding is dependent upon accurate commodity coding. At present the program acts only as an internal consistency check--with one very frequent outcome

being corrected commodity codes. An alternative means of improving commodity coding would be to use questionnaires which have been personalized to show respondents the official descriptions corresponding to the previous year's response.

(iv) Industry classification is also dependent upon accurate shipment values. Aside from the usual handling errors, there is a conceptual issue. In the Canadian Census of Manufactures, commodity data are intended to relate to value shipped net of: sales and excise taxes, discounts, returns and transportation charges. Sometimes there is only an aggregate, plant-level, figure for such impurities. Adjustments would almost certainly originate at different rates between different products. The unadjusted product data are used. The impact of such distortions is presumed to be small.

(v) There is a lack of annual commodity data for progress payments industries--those are industries in which a product is only delivered after several years of work. Detailed product data in any given year may not reflect the actual level of activity in the plant. Typically, these plants are too few and too large to miss.

(b) Classification-Related Considerations

(i) Manufacturing services are treated as primary where reported. Generally referred to as custom and repair work, such activity comprises services performed by a manufacturing unit on goods not owned by itself. [9] Since Statistics Canada has no official services classification, this is covered by codes, supplementary to the Industrial Commodity Classification, called pseudo-ICCs. There are only a few

such codes and because of their generality there are not many cases where a unique primary industry affiliation exists.

Ignoring custom and repair work when verifying industry assignments would lead to rejection of valid industry designations and would burden industry officers and editors with unnecessary work. In this application, treating custom and repair work as primary to the reporting establishment's assigned industry is practical because an industry assignment has already been made. Such treatment is not feasible when engaged in automated coding where there is no such previous assignment. The solution is to adopt a detailed services classification which can be uniquely linked to four-digit industry classes.

(ii) On occasion, gaps can appear in the commodity classification system--it may handle only in a general way new products or those for which the overall value is low. This is a problem for this routine only if the primary industry assignment is inappropriate. If additional commodity classes are introduced, primary links will have to be established with the current industry classification--as well as with any previous systems for which extended time-series might be generated through automated classification of individual units. In addition, once more detailed commodity classes are established, those less detailed levels for which an aggregation problem exists should be eliminated from use.

(iii) Industries may be defined by the process of production rather than by the products per se which can, in fact, show considerable year-to-year variation. Cases of joint production and of vertically integrated production were discussed above. However, it may not always be possible to capture process dimensions in commodity descriptions.

The Machine Shop Industry, SIC 3081, involves such difficulties--along with a considerable amount of custom and repair work. When engaged in production on own account, members of this industry may manufacture a variety of goods primary to other industries. This leads to frequent error messages. At present, a plant which specializes for a period of time in the products of some other industry will be transferred to that other industry. For example, a unit making only eating utensils for several years would be moved to the cutlery industry. There is a possibility that such transfers could involve bona fide machine shops. It is occasionally suggested that the membership of the industry simply be frozen--which would bypass the edit. The real solution is to ask process-oriented questions--or to avoid establishing industry classes based on such criteria.

(c) Other Considerations

(i) Sequential editing and release of industry data mean that it may not be possible to transfer establishments examined at later stages of that process. Dealing with this limitation would require a fundamental change to the processing system.

(ii) It can be the case that two or more industry designations are calculated as equally applicable. This type of situation usually results from the respondent or editor having estimated individual outputs by using round percentages to distribute the total. This is most likely to occur on short forms where percentages rather than dollar values are asked.

(iii) The application of resistance rules is presently somewhat subjective. There is a broad notion that an establishment which is just marginally into an industry but remains so for two or three years, should be transferred. Such rules of thumb could be codified.

(iv) On occasion, coverage is allowed to supercede specialization as a determinant of industry coding. This is thought to be one of the benefits of manual classification, but the general principles involved could be made explicit and become part of the series of prespecified rules required for automated coding.

Conclusions

Use of the Questionably-Coded methodology has led to considerable data improvements in the Census of Manufactures--both in terms of industry and commodity coding. With the adoption of a detailed classification of manufacturing services, and with increased codification of existing practices, the number of manual interventions can be decreased and greater standardization can be achieved in the implementation of the classification. The edit can accommodate data collected under the Harmonized Commodity Description and Coding System, and it can be extended to industry classification in other divisions where detailed product data are collected. More generally, procedures of this nature might be of use in any survey involving classifications that can be related to one another. For example, it is conceivable that one could devise a similar edit to verify occupational classification.

Notes

1. Jorge Luis Borges, Other Inquisitions: 1937-1952 cited in Mark S. Aldenderfer and Roger K. Blashfield, Cluster Analysis. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-044. Beverly Hills and London: Sage Publications.

2. See for example Andrews & Abbott.

3. The links to the 1980 Standard Industrial Classification are published in Manufacturing Industries of Canada: national and provincial areas, 1983, Statistics Canada, Catalogue 31-203. A list of products primary to each 1970-based industry is given in Concepts and definitions of the census of manufactures, 1979, Statistics Canada, Catalogue 31-528. These links are derived conceptually rather than empirically and, in fact, there are many commodity classes where the majority of shipments is secondary. Use of primary products to classify establishments to industries, by the 1947 U.S. Census of Manufactures, is discussed in Conklin & Goldstein.

4. Short forms are less detailed questionnaires generally sent to births and to smaller establishments.

5. All the manufacturing data in this paper include logging and forestry.

6. This compares successive years' classification for the establishments covered by this edit. It thereby nets out cases where transfers were undone, and counts only

once cases that were retransferred (to a third industry) during the processing of the data for a given year.

7. See Potter for an extended discussion of the nature and certain uses of these measures.

8. Similar refers to the industry's defining activities, but even these can involve a considerable degree of conceptual dissimilarity. Consequently, any two plants within an industry may be quite different. To the extent that there are separate clusters of conceptually similar plants within an industry, sub-industries exist. This would likely apply to the residual classes which comprise a number of small but conceptually distinct industries.

9. In 1985, some 6,385 establishments reported custom and repair work having a total value of $C 6.2 billion.

References

Stephen H. Andrews and Thomas A. Abbott III, 'An examination of the standard industrial classification of manufacturing activity using the longitudinal research data base' in Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census (Washington, D.C., 1988) 467-488.

Maxwell R. Conklin and Harold T. Goldstein, 'Census principles of industry and product classification, manufacturing industries' in National Bureau of Economic Research Conference Report, Business Concentration and Price Policy, (Princeton University Press, Princeton, 1955), 15-36.

James W. McKie, 'Industry classification and sector measures of industrial production,' U.S. Bureau of the Census Working Paper No. 20, (Washington, D.C., 1965).

Harley Potter, 'Some conceptual aspects of measuring homogeneity of industrial data from manufacturing censuses and surveys' Statistical Journal of the U.N. Economic Commission for Europe, (Volume 5, No 4., 1988).

Statistics Canada, Concepts and definitions of the census of manufactures, (Catalogue 31-528 Occasional, Ottawa, 1979).

Statistics Canada, 'Notes on the 1980 Standard Industrial Classification in the manufacturing industries' in Manufacturing Industries of Canada: national and provincial areas, 1983, (Catalogue 31-203 Annual, Ottawa, 1986), xxiii-xcviii.

Statistics Canada, Standard Industrial Classification 1980, (Catalogue 12-501E, Ottawa, 1980).

Statistics Canada, Standard Industrial Classification Manual, Revised 1970, (Catalogue 12-501 Occasional, Ottawa, 1970).

U.S. Bureau of the Census, 1977 Census of Manufactures, Volume 1, Subject Statistics, (Washington, D.C., 1981), x-xiv.

Appendix

Questionable SIC Coding in Year 1985: SIC 1999 Other Textile Products Industries NEC
ICC-SIC Project, Analysis Section, Industry Division.
John Crysdale

This RSN is in Status 7

This is RSN: 1234569

Reason Code: _____

No SIC Override yet this yr
Last year's SIC Override: 2
No SIC Override prev year.

(A) I am changing this SIC to _____. If this is to take effect next year, also circle 3 or 5.

(B) I am keeping the present SIC--but the ICCs or their values are being amended, or a PSIC is being changed. The problem should disappear by next run.

(C) I am keeping the present SIC and need an exception code. Put reason number in box. Return to Crysdale.

(1) ICC related - New ICC classes needed. Most detailed available ICC covers a range of products and PSICs. Where this represents significant value, I will request new ICCs be broken out.
- PSIC=0: more precise commodity specification needed for automated SIC calculation. (More detailed ICC classes do exist and are primary to a range of SICs.) I cannot get a more precise breakdown and I consider this RSN to be in the proper SIC. Where this is a pre-printed ICC, next year's questionnaire will be modified.

(2) Stability (Resistance Factor): current year atypical. Includes continual shifting. Will monitor.
(3) Would transfer but this or recipient SIC frozen (or very close to it). Will transfer for next year.
(4) Other. Explain.
(5) These values will not have a significant effect on SIC data. I am not taking further action this year.

Signature: _____

1999-5-1234569-Bonnie Days Outdoors Co (East Plant) Form=L QSIC=1931 Spec ratio in current SIC: 35 / in QSIC: 45

Unadj PSIC	PSIC	ICC	ICC Name	Total Value ($ 000s)	Value Primary to Actual SIC ($ 000s)	Value Not Primary to Actual SIC ($ 000s)
1999	1999	965 4	Sleeping bags	2,500	2,500	0
1999	1999	965 51	Flags (incl penants), textile	1,000	1,000	0
1931	1931	965 312	Tents, hikers' & children's	4,000	0	4,000
1931	1931	747 32	Awnings, canvas	500	0	500
1699	1699	747 33	Awnings, plastic	1,750	0	1,750
1999	0	965	Textile end products	250	0	250
Data year 1985, RSN in SIC 1999, Totals:				10,000	3,500	6,500
1999	1999	965 4	Sleeping bags	2,250	2,250	0
1999	1999	965 51	Flags (incl penants), textile	1,000	1,000	0
1931	1931	965 312	Tents, hikers' & children's	3,500	0	3,500
Data year 1984, RSN in SIC 1999, Totals:				6,750	3,250	3,500
1999	1999	965 4	Sleeping bags	2,000	2,000	0
1999	1999	965 51	Flags (incl penants), textile	2,000	2,000	0
Data year 1983, RSN in SIC 1999, Totals:				4,000	4,000	0

ANALYTICAL STUDIES BRANCH

RESEARCH PAPER SERIES

No.

1. BEHAVIOURAL RESPONSE IN THE CONTEXT OF SOCIO-ECONOMIC MICROANALYTIC SIMULATION by Lars Osberg

2. UNEMPLOYMENT AND TRAINING by Garnett Picot

3. HOMEMAKER PENSIONS AND LIFETIME REDISTRIBUTION by Michael Wolfson

4. MODELLING THE LIFETIME EMPLOYMENT PATTERNS OF CANADIANS by Garnett Picot

5. JOB LOSS AND LABOUR MARKET ADJUSTMENT IN THE CANADIAN ECONOMY by Garnett Picot and Ted Wannell

6. A SYSTEM OF HEALTH STATISTICS: TOWARD A NEW CONCEPTUAL FRAMEWORK FOR INTEGRATING HEALTH DATA by Michael Wolfson

7. A PROTOTYPE MICRO-MACRO LINK FOR THE CANADIAN HOUSEHOLD SECTOR SESSION 3, MACRO/MICRO LINKAGES - HOUSEHOLDS by Hans Adler and Michael Wolfson

8. NOTES ON CORPORATE CONCENTRATION AND CANADA'S INCOME TAX by Michael Wolfson

9. THE EXPANDING MIDDLE: SOME CANADIAN EVIDENCE ON THE DESKILLING DEBATE by John Myles

10. THE RISE OF THE CONGLOMERATE ECONOMY by Jorge Niosi

11. ENERGY ANALYSIS OF CANADIAN EXTERNAL TRADE: 1971 and 1976 by K.E. Hamilton

12. NET AND GROSS RATES OF LAND CONCENTRATION by Ray Bollman and Philip Ehrensaft

13. CAUSE-DELETED LIFE TABLES FOR CANADA (1921 to 1981): AN APPROACH TOWARDS ANALYSING EPIDEMIOLOGIC TRANSITION by Drhuva Nagnur and Michael Nagrodski

14. THE DISTRIBUTION OF THE FREQUENCY OF OCCURRENCE OF NUCLEOTIDE SUBSEQUENCES BASED ON THEIR OVERLAP CAPABILITY by Jane F. Gentleman and Ronald C. Mullin

15. IMMIGRATION AND THE ETHNOLINGUISTIC CHARACTER OF CANADA AND QUEBEC by Réjean Lachapelle

16. INTEGRATION OF CANADIAN FARM AND OFF-FARM MARKETS AND THE OFF-FARM WORK OF WOMEN, MEN AND CHILDREN by Ray Bollman and Pamela Smith

17. WAGES AND JOBS IN THE 1980s: CHANGING YOUTH WAGES AND THE DECLINING MIDDLE by J. Myles, G. Picot and T. Wannell

18. A PROFILE OF FARMERS WITH COMPUTERS by Ray D. Bollman

19. MORTALITY RISK DISTRIBUTIONS: A LIFE TABLE ANALYSIS by Geoff Rowe

20. INDUSTRIAL CLASSIFICATION IN THE CANADIAN CENSUS OF MANUFACTURES: AUTOMATED VERIFICATION USING PRODUCT DATA by John S. Crysdale

21. CONSUMPTION, INCOME AND RETIREMENT by A.L. Robb and J.B. Burbridge

For further information, contact the Chairperson, Publication Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A 0T6.